

材料辞書データベースを使った論文からの大量データ抽出：
用語間関係性抽出の自動化検討

Automatic data extraction from scientific articles using the materials-dictionary database:

Development for automatic entity linking

物材料機構 ○鈴木 晃, 石井 真史

National Institute for Materials Science, °Akira Suzuki, Masashi Ishii

E-mail: SUZUKI.Akira3@nims.go.jp

1. 背景

本研究では、マテリアルズ・インフォマティクス(MI)用学習データを効率的に収集するための技術を構築している。これまで、大量の学術論文から抽出した材料用語を自動階層構造化することで材料辞書データベース (DB) を構築し、論文内用語に対し体系的に自動タグ付けする手法を開発してきた[1]。MI 用学習データ取得のためには、タグ付けされた用語間 (材料名、実験手法、物性名、物性値等) の紐づけ (関係性抽出) が必要となる。本稿では関係性抽出を自動化するため、まずは上記タグの中から 1 対のタグ間の関係性抽出を行った。また、略語とその正式名称の紐づけも追加し材料辞書 DB を拡張した。

2. 材料辞書 DB の拡張

2.1 用語間の関係性抽出

材料辞書 DB は学術論文から抽出した材料用語を論文分類や物性データ抽出等の目的に応じて階層分類した RDF (Resource Description Framework) 形式のデータベースである。本研究では拡散関連 (“diffusion coefficient”、“activation energy”が含まれる) 論文を対象としている。論文から抽出した用語 (複合語) に対して複合語の最後の単語 (以下、Base word) を基にした階層分類を行うことで半自動的 (人手による修正を含む) に用語にタグ付けする手法を確立した。本稿ではこれに加え、複合語の最初の単語を Base word とした階層分類を試みたところ、“diffusion coefficient of ...”、“analyzed by ...”といった複合語による分類(Base phrase)構造となった。それぞれの複合語は、既存の材料辞書 DB により前置詞前後の用語が属するタグを参照することにより “property-material”、“method-equipment”といった関係性を付与することができ、Table 1 に例示する材料辞書 DB が構築された。

2.2 略語辞書の追加

学術論文では略語が多用され、初出時は “scanning electron microscopy (SEM)” のように正式名称と略語が記載されるものの、以降は略語のみである。よって材料辞書内の用語も略語を含む複合語となる (例えば Table 1 内の Base phrase : analyzed by)。そこで、対象論文から略語と正式名称の対を抽出し、材料辞書 DB に登録することで正式名称を参照できるようにした。

3. まとめ

材料用語間の関係性抽出自動化を目的に、1 対のタグ間の関係性抽出、略語と正式名称の紐づけを行うことで材料辞書 DB の拡張を行った。MI 用学習データとして使用するためには、さらに複数タグ間の関係性を抽出する必要があり、現在、検討を進めているところである。

【謝辞】

本研究の一部は、文部科学省「元素戦略磁性材料研究拠点」委託業務の一環として実施されました。

参考文献

[1] 鈴木 晃, 石井 真史, 第 81 回応用物理学会秋季学術講演会, 9p-Z09-17(2020).

Table 1 Examples of materials dictionary

Base phrase	Phrases*	Specification
diffusion coefficient of	diffusion coefficient of alkali ion, diffusion coefficient of anion and cation, diffusion coefficient of Ca2+, diffusion coefficient of carbon dioxide, diffusion coefficient of CO and CH, diffusion coefficient of component i, diffusion coefficient of Cu Sn, diffusion coefficient of DCM and TCM, diffusion coefficient of Li ion	property-material
	diffusion coefficient of 0 × 0 10, diffusion coefficient of 0 × 0 12, diffusion coefficient of 0 × 0 15, diffusion coefficient of 0 × 0 6, diffusion coefficient of 0 × 0 7	property-value
annealed at	annealed at 0 K, annealed at 0 K for 0 h, annealed at 0 ° C, annealed at 0 ° C for 0 h, annealed at 0 ° C in Ar atmosphere, annealed at 0 ° C in vacuum, annealed at different condition, annealed at different temperature	process-condition
analyzed by	analyzed by energy dispersive X ray spectroscopy, analyzed by field emission scanning electron microscopy, analyzed by GC MS, analyzed by SEM, analyzed by SIMS, analyzed by TEM, analyzed by TGA, analyzed by XPS, analyzed by XRD	method-equipment
concentration dependence of	concentration dependence of activation energy, concentration dependence of chemical diffusion coefficient, concentration dependence of diffusion coefficient, concentration dependence of self diffusion coefficient, concentration dependence of viscosity	condition-property

*Numbers are normalized with “0”